RAND

The Performance of Students With Disabilities on New York's Revised Regents Comprehensive Examination in English

Daniel Koretz, Laura Hamilton

DRU-2608-EDU

July 2001

Prepared for UCLA's National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

RAND Education

The Performance of Students With Disabilities on New York's Revised Regents Comprehensive Examination in English

CSE Technical Report 540

Daniel Koretz
CRESST/RAND Education
Laura Hamilton
RAND Education

April 2001

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.3 Accommodation, Daniel Koretz, Project Director, CRESST/RAND Education

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

THE PERFORMANCE OF STUDENTS WITH DISABILITIES ON NEW YORK'S REVISED REGENTS COMPREHENSIVE EXAMINATION IN ENGLISH[1]

Daniel Koretz, CRESST/RAND Education
Laura Hamilton, RAND Education

Abstract

Federal and state policy initiatives are greatly expanding the inclusion of students with disabilities in large-scale assessments, but there is little experience or research to guide this effort. Earlier CRESST studies (Koretz, 1997; Koretz & Hamilton, 1999) examined the experience of Kentucky, one of the first states to include the large majority of students with disabilities in its assessment. The studies revealed a number of important technical and practical issues. State assessments differ markedly, and experience with inclusion may vary from state to state. Accordingly, this study explored the performance of students with disabilities in a field test of the revised New York State Regents Comprehensive Examination in English, the first of the new Regents examinations that almost all students in that state will have to take to obtain a high school diploma. Data from the field test were gathered statewide but not necessarily from a fully representative sample of schools. Accommodations were used liberally, with extra time and testing in a separate location being the most common. Completion rates were similar for students with and without disabilities, and few items had very low p values for students with disabilities. However, students with disabilities scored roughly two thirds to one and one third standard deviations below other students, and a high percentage of students with disabilities provided either unscorable or extremely weak responses to open-response items. The study clearly underscores the need for more extensive information to clarify the effects of including students with disabilities in high-stakes assessments. In addition, it raises concerns about possibly excessive levels of difficulty for some students with disabilities, which could cause either very high failure rates or undesirable responses by teachers or students, such as excessive coaching.

Over the past several years, extensive efforts have been made to include students with special needs in the large-scale assessments administered to the general education population. Until the mid-1990s, the inclusion of such students was inconsistent, and substantial percentages of them often were excluded.

[1] We are grateful to several people for assistance with this work. First, we thank James Kadamus, deputy commissioner of the New York State Education Department, and Gerald DeMauro, director of the Office of State Assessment, for access to the data used in this report. We want to thank Dr. DeMauro, Karen Kolanowski, and Tom Schoeck of the Education Department for helpful reviews of a draft of this report as well as many other kinds of assistance during the course of the project. Helpful comments also were provided by an anonymous CRESST reviewer. Chi San and Robin Beckman assisted with statistical programming, and Christel Osborn formatted the document. We remain solely responsible, however, for any errors of fact or interpretation.

However, the Improving America's Schools Act of 1994 amended Title I (the principal federal compensatory education program) to require that Title I students be assessed with the same tests used for other students and that scores be reported separately for several categories of students with special needs, including students with disabilities. In 1997, amendments to the Individuals with Disabilities Education Act (IDEA) required that students with disabilities be included in state and district assessments to the extent feasible and that alternative assessments be administered to those few students unable to participate in the regular assessment. Policy initiatives in numerous states, such as Kentucky and Maryland, showed the same impetus to increase and regulate the inclusion of students with disabilities in large-scale assessments. The goal of this inclusion is to make schools more accountable for the achievement of students with disabilities and to encourage their greater incorporation into the general education curriculum.

The movement toward greater inclusion of students with disabilities was concurrent with the continued strengthening of the standards-based reform movement. This entailed major changes in the nature of large-scale assessments, often including test questions of greater difficulty, more reliance on extensive reading and writing in all subject areas, the use of performance tasks, and the deliberate mixing of skills and types of knowledge. Moreover, the results of these assessments often are reported only in terms of the percentages of students meeting a small number of standards (often three), and the lowest of these often is high relative to the current performance of many students with disabilities.

The goal of increasing the inclusion of students with disabilities in these new assessments is hindered by a dearth of relevant research and experience (National Research Council, 1997). For example, research on the effects of K-12 assessment accommodations is sparse. Similarly, little is known about the impact of the characteristics of the new assessments—such as their reliance on written responses—on the performance of students with disabilities.

In response to this lack of information, CRESST has undertaken several studies of the assessment of students with disabilities. The first (Koretz, 1997) examined the performance of students with disabilities on the statewide assessment (KIRIS) in Kentucky, which was in the vanguard of increased inclusion. A follow-up study by Koretz and Hamilton (1999) replicated and extended that study using newer data that allowed a direct comparison of performance on multiple-choice (MC) and constructed-response items. These studies found that the large majority of identified

students with disabilities were included in the main Kentucky assessment. Accommodations were used extensively, particularly in the lower grades. These had more of an effect on open-response (OR) questions, and in some instances, students with accommodations had implausibly high scores. Differential item functioning, a sign of possible bias, was found in both test formats, but only among students who received accommodations.

This study conducts similar analyses of the performance of students with disabilities in the first field test of New York State's revised Regents Comprehensive Examination in English. The Regents English test is the first of the new Regents examinations to be required of all high school graduates. It offers a potentially informative contrast to Kentucky's assessment in several respects. First, the Regents English test, unlike the Kentucky assessment, has high stakes for individual students, and the resulting motivational factors may influence differences in performance between students with and without disabilities. Second, the Regents English examination is a different type of test, including a listening task and stressing even more than the KIRIS assessment the writing of extended responses. Third, the Regents English test is used in a different context, with a very different tradition of state testing (dating back well over 100 years) and different demographics. New York and Kentucky also differ in their policies concerning assessment accommodations. In New York, for example, extra time—the most common accommodation in many programs—is allowed only for students who are specifically entitled to that as an accommodation, e.g., because of an Individualized Education Program (IEP), and extra time is recorded as an accommodation. In Kentucky, all students who wanted extra time to complete the KIRIS assessment were offered it, and the provision of extra time was never recorded.

Description of Assessment and Data

The new Regents English examination comprises four parts. In the operational assessments, students are administered all four parts. The operational examination is administered in two 3-hour sessions; students complete Parts I and II in the first session and Parts III and IV in the second session. The New York State Education Department (NYSED) documents both expected and "allocated" times, but it does not intend that students be stopped at any time during the three-hour sessions to advance to the second part of the exam.

Part I, "Listening and Writing for Information and Understanding," is intended to discern how well students can "interpret and analyze complex information presented orally" (New York State Education Department, n.d. b). Students are required to listen to a passage read aloud, answer MC questions pertaining to the passage, and then write an extended response. NYSED refers to these as "listening comprehension" prompts, but we use the more descriptive term "read-aloud" to avoid assumptions about the skills measured by Part I.

Part II, "Reading and Writing for Information and Understanding," is intended to measure how well students can interpret and analyze information from written text and graphics. This part is similar in demands to Part I except for the nature and presentation of the passage; it too entails both MC questions and a single extended response. Expected time is 90 minutes; allocated time in the absence of an IEP or 504 plan is 105 minutes.

Part III, "Reading and Writing for Literary Response and Expression," focuses on the ability to interpret literary texts. Students are required to read two texts, answer MC questions, and write one extended response. Expected time is 90 minutes; allocated time in the absence of an IEP or 504 plan is 105 minutes.

Part IV, "Reading and Writing for Critical Analysis," focuses on the analysis of literary texts. This part has a different structure from the other three. Students are presented with a statement or quotation about literature. They are asked to explain it, state their opinion about it, select two literary works they have read that support their opinion, and discuss specific elements of these works in developing support for their arguments. Expected time is 45 minutes; allocated time in the absence of an IEP or 504 plan is 75 minutes.

The extended responses were scored using an approach that has been called in other contexts a "focused holistic" approach. The scoring rubrics for the four parts differed in some details but were generally similar. Each listed five "qualities" or aspects of performance: meaning, development, organization, language use, and conventions. The rubrics described in general terms what was required to reach each of six score points. For example, for Part I, the rubric described a rating of 3 on development as "a response...[that] develops ideas briefly, using some details from the text," while responses at the higher score of 4 "develop some ideas more fully than others, using specific and relevant details from the text." Although score points

are defined for all five qualities, raters provide a single score for the extended response, rather than individual scores on each individual quality.

The data analyzed here were generated by a field test of the new Regents English test administered in October 1998. For present purposes, field test data suffer from several weaknesses noted later, but they are particularly important in the New York Regents system because they are used to set the scale that determines whether students pass. Data from an operational administration of the new assessment were not available for comparison.

Although the Regents English test will most often be administered to students completing their junior year, the field test was administered to students at the beginning of their senior year because of the test development schedule. Three samples of schools were drawn, each of which received one type of test form (described later). We refer to the three as "subsamples" to avoid confusion with the total field test sample. The sample frame was created by dividing all schools in the state into four categories depending on the percentage of students who passed the Regents English Examination—the predecessor to the revised Regents English test. NYSED (n.d. a) described the sampling as follows:

> Schools were assigned to each of the three [sub]samples so that within each sample there was an equal number of students representing each category of performance. Each [sub]sample also contained schools representing a variety of community types and geographic regions.

We were not able to obtain information, however, about participation rates or whether the final sample was representative of the participating schools. In addition, no information was collected about the exclusion of students with disabilities from the field test. This hinders interpretation of the results of this analysis, as will be described.

The field test included six variants of each of the four parts of the Regents English test. Although the operational forms of the test require that students take all four parts, most students in the field test were administered only two parts. These parts were combined into 16 forms that included two parts and two additional forms that included four parts. Because the four-part forms were administered to few students and were not comparable to the others, we included only the 16 two-part forms in our analysis. All of these 16 forms included two extended-response tasks. Because the parts differed in structure, however, these 16 forms included

varying numbers of MC items: 6, 10, 16, or 20. These forms were administered in a single 3-hour sitting, except to the extent that accommodations were offered, as noted below.

Each of the three subsamples was administered one type of assessment. The first subsample was administered forms with one listening task and one reading task. The second subsample was administered forms with two reading tasks. The third, smaller subsample, which we excluded from our analysis, was administered the two forms with all four types of tasks.

Participating schools provided NYSED with class rosters. All 12th-grade students on these rosters were assigned randomly to forms by NYSED. Students who were assigned the same form were tested together. Staff members were asked not to inform students about the link between the field test and the new Regents English examination.

Information on sampling and participation is limited. We were provided the following information about counts (K. Kolanowski, personal communication, July 24, 2000):

> Number of Students Contacted—1,200 per full [four-part] form...; 2,500 per other [two-part] forms...;
>
> Number Sent to Schools—400 per full form; 1,200-1,360 per other forms;
>
> Number Administered (Usable)—237-278 per full form; 554-670 per other forms.

The number of students actually sent to schools was much smaller than the number contacted. Some principals refused to participate entirely, while a few agreed to administer the test but only to a smaller number of students. In the latter case, principals were instructed to insure the smaller number of students was representative (K. Kolanowski, personal communication, July 27, 2000). The number sent to schools and the "number administered (usable)" dropped substantially for numerous reasons: principals' enrollment estimates often were inflated or rounded up; some students inevitably were absent; and in some schools, absences increased because students knew the field test would not count for them (K. Kolanowski, personal communication, July 27, 2000). In addition, students who lacked either an MC answer sheet or a scored OR record were dropped from NYSED's file of scored records, as were students who had left either the MC or open-ended section blank. However, the data file given to us included numerous students in the data set who

did not in fact complete all four parts (as described in a later section). The data sent to us included test scores for 12,555 12th-grade students, although records for some of the students were incomplete.

Some of these participation problems would not bias the sample (e.g., a principal's willingness to administer the test only to a random subset of the eligible students), but many of them would. The non-participation rates are high enough to pose a potentially serious threat to the representativeness of the field test sample. We had little information that would allow us to explore the characteristics of the participants and non-participants and the representativeness of the final sample.

Letters to principals requested that "all grade 12 students" be tested. Principals in participating schools were reminded that students may be entitled to accommodations because of IEPs or 504 plans but were given no additional instructions about this. Tabulations of the data described later suggest some exclusion of students with disabilities, but no records of exclusions were kept.

NYSED allows testing accommodations, including extra time, under four circumstances.[2] Students entitled to accommodations are:

1. students with IEPs that call for accommodations;
2. students who recently have been declassified—that is, determined not to need further placement in special education—whose declassification documents a need for continuation of accommodations specified in the IEP;
3. students with disabilities who have a Section 504 Accommodation Plan that includes test modifications; and
4. STUDENTS who have been classified as disabled shortly before test administration, including students with temporary disabilities. (New York State Education Department, 1995)

The data included records of accommodations offered to students with disabilities. However, proctors noted the use of an accommodation if it was used at any point in the exam, and it is not possible to determine whether they used accommodations differently on the MC and OR parts of the test.

---

[2] NYSED refers to these as "test modifications," defined as "changes in testing procedures or formats" (Office of Vocational and Educational Services for Individuals with Disabilities, 1995). The terms "modification" and "accommodation" are used inconsistently across testing programs, but changes in presentation or mode of response are more often labeled as accommodations.

As noted, the data supplied to us included test scores for 12,555 12th-grade students. Records for approximately 30% of these (3,805 students), however, included no information on disability status. Most of these students came from schools where no disability information was reported for any student. Because we have no way of knowing how these students should be classified, we eliminated them from our analyses.

Prevalence of Disabilities and Accommodations

Of the 8,750 students in the sample with disability data, 563 (6.4%) were classified as having at least one disability. This is markedly lower than the percentage of New York students who are classified as disabled. In the 1996-'97 school year, approximately 12% of New York children ages 6-17 who were enrolled in school were served under IDEA, Part B (U.S. Department of Education, 1998, Table AA12, p. A-37), and some additional children not served under IDEA were presumably identified as disabled under Section 504. While the percentage of high school students served by IDEA is often lower than the percentage of younger students, it is likely that the field test sample included a substantially smaller percentage of students with disabilities than would be found in the entire age group statewide. This suggests either that schools excluded a substantial proportion of identified students with disabilities from the field test or that the schools that participated in the field test were atypical in terms of their identification rates.

It is not clear, however, whether the rate of exclusion of students with disabilities will be similar in the operational administrations of the Regents English examination. There were no consequences for participation in or performance on the field test, which may have led to a higher exclusion of students who were not likely to do well on the test. Also, because the new Regents examinations are not tied to a specific grade or age, the exclusion rate will not be apparent from data from a single operational administration of the examination. Rather, it will be necessary to accumulate data over a number of years to discern the percentage of each cohort of students with disabilities that takes the Regents English examination.

Data for all 563 students with disabilities were used to describe the sample and the testing accommodations used for them. However, only students who had scores for both the MC and OR portions of the assessment were included in tabulations that involved performance, in order to avoid confounding differences in performance with differences between subsamples. Only 481 students with

disabilities, 85% of all sampled students with disabilities, had scores for both parts of the assessment.

The small number of students with disabilities severely hampered analysis of these data. The limitation of small sample size was compounded by the non-equivalence of forms (because the number of students with disabilities who were administered any single form was very small) and the heterogeneity of students with disabilities. Because of the small sample, only very large differences in performance would reach conventional levels of statistical significance. Accordingly, we present most results here without significance tests. These findings are merely descriptive and suggestive; in many cases, additional data would be needed to determine how much confidence to place in them.

Disabilities of Tested Students

Table 1 shows that nearly 80% of students with disabilities were classified as having learning disabilities. This is a substantially higher percentage than in New York State as a whole. In the 1996-1997 school year, approximately 65% of the students of secondary age (12-17 years) served under IDEA, Part B, in New York were identified as having specific learning disabilities (U.S. Department of Education, 1998, Table AA4, p. A-8). This difference could indicate either that the sample of schools participating in the field test was unrepresentative or that students with other disabilities were excluded from the field test at a higher rate than students with learning disabilities.

Only a modest number of students were classified as having more than one disability. About 87% of the students with disabilities were classified as having a single disability (Table 2). Eight percent were assigned to the category "multiple disabilities," with no further information about the number or type of disabilities. Four percent were classified as having two specific disabilities, and a small number were classified as having more than two.

Given the heterogeneity of students with disabilities, it would be preferable to analyze the performance of the more homogeneous groups of students sharing a disability classification. For example, a given type of test item might be biased for students with visual disabilities but not for students with learning disabilities. Unfortunately, the numbers of students with disabilities in our sample made this impossible for most disability groups. It was, however, feasible to conduct many analyses separately for students with learning disabilities, who constituted the

Table 1

Frequencies and Prevalence of Specific Disabilities

Disability	*N*	Percent of disabled sample
Learning disabled	446	79.2
Multiple disabilities	45	8.0
Hearing impaired	35	6.2
Other health impaired	33	5.9
Visually impaired	23	4.1
Emotionally disturbed	22	3.9
Orthopedic impaired	14	2.5
Speech impaired	11	2.0
Mentally retarded	9	1.6
Hard of hearing	7	1.2
Autistic	5	0.8
Deaf/blind	3	0.5

Note. Several students were assigned more than one of the other categories without being classified under "multiple disabilities." Consequently, percents sum to more than 100.

Table 2

Numbers of Disability Categories Assigned to Students With Disabilities

Number of disability categories	*N*	Percent
1	491	87.2
2	23	4.1
3	2	0.4
4	1	0.2
6	1	0.2
Single listing of "multiple disabilities"	45	8.0
Total	563	100

majority of students with disabilities in our sample. The analyses of students with learning disabilities paralleled those for all students with disabilities and are generally reported after them. Analyses reported here reflect our entire sample of students with disabilities unless otherwise noted.

Use of Accommodations

Testing accommodations were liberally used. Nearly three fourths of all students with disabilities were given one or more testing accommodations (Table 3). More than half of the students were given extended time, and more than half were tested in a separate location. Directions were read or clarified for about 33% of students with disabilities, and the test was read aloud to 30% of them. Relatively few students received any of the other recorded accommodations.

The rate of use of accommodations was slightly higher for students with learning disabilities than for all students with disabilities (Table 4). Note that the findings for all students with disabilities are largely shaped by the results for students with learning disabilities, who constituted nearly 80% of the sample of students with disabilities. The small number of students with other disabilities precludes tabulating them separately. It is useful to examine students with learning disabilities separately, however, because they represent a particularly large and more homogeneous group.

Table 3

Percent of Students With Disabilities Receiving Accommodations

	All students with disabilities		Students with learning disabilities	
Accommodation	*N*	%	*N*	%
Any accommodation	404	71.8	346	77.6
Any accommodation other than time extension or separate location	266	47.2	223	50.0
Time extension	308	54.7	271	60.8
Separate location	307	54.5	263	59.0
Directions read/clarified	183	32.5	152	34.1
Test read to student	167	29.7	144	32.2
Spell checker and/or grammar checker	62	11.0	55	12.3
Spelling/punctuation/paragraph waiver	33	5.9	29	6.5
Amanuensis/scribe/tape recorder	11	2.0	8	1.8
Other accommodations	61	10.8	48	10.8

More than two thirds of the accommodated students—just over half of all tested students with disabilities—were given more than one accommodation. About 29% of students receiving accommodations, corresponding to 21% of all tested students with disabilities, received only a single accommodation (Table 4). Fifty-three percent of students who received any accommodation, corresponding to 38% of all tested students with disabilities, received three or more accommodations.

Although multiple accommodations were much more common than single accommodations, no specific combinations of accommodations were used with as many as 10% of students with disabilities. The two most common combinations of accommodations were time extension and separate location (9% of tested students with disabilities) and time extension, separate location, test read, and directions read or clarified (8%; see Table 5). The pattern of use of accommodations was essentially the same for students with learning disabilities as for all students with disabilities.

These patterns in the use of accommodations can be compared to those found in Kentucky's KIRIS assessment program, one of the first in the nation to include the great majority of students with disabilities. By the mid-1990s, KIRIS included 85% to 90% of 11th-grade students with disabilities (Koretz, 1997). The two accommodations most commonly used in New York, additional time and a separate location, must be excluded when comparing New York to Kentucky because these two accommodations were not identified in the Kentucky data. Additional time was available to any Kentucky student who needed it, whether disabled or not, and

Table 4

Numbers of Accommodations Given

Number of accommodations	*N*	Percent of students with disabilities	Percent of accommodated students
0	159	28.2	
1	118	21.0	29.2
2	72	12.8	17.8
3	63	11.2	15.6
4	95	16.9	23.5
5	37	6.6	9.2
6	17	3.0	4.2
7	2	0.4	0.6

Table 5

Most Frequent Combinations of Accommodations

Accommodation	All students with disabilities		Students with learning disabilities	
	N	%	*N*	%
Time extension only	58	10.3	52	11.7
Separate location only	28	5.0	23	5.2
Time extension and separate location	52	9.2	48	10.8
Time extension, separate location, test read, and directions read or clarified	46	8.2	38	8.5
Time extension, separate location, and directions read or clarified	26	4.6	23	5.2
Time extension, separate location, and test read	25	4.4	23	5.2

teachers were not asked to indicate on the test forms whether additional time had been provided. The forms used in Kentucky also did not ask teachers to indicate whether the assessment had been administered in a separate location.

The use of accommodations other than extra time and a separate location was somewhat less common in New York than in Kentucky. Similarly, the use of multiple accommodations was less common in New York than in Kentucky. In New York, 47% of students with disabilities received at least one such accommodation; about 19% received one, and about 29% received more than one (Table 6). In contrast, in Kentucky, 61% of students received at least one such accommodation; 23% received one, and 39% received more than one.

This comparison between New York and Kentucky could be distorted by at least two factors: (a) the apparently greater exclusion from the assessment of students with disabilities other than learning disabilities in New York, and (b) the different lists of accommodations allowed and tabulated in the two states. To address the first of these, we repeated these analyses including only students with learning disabilities in order to obtain groups that are presumably more comparable between the two states. The results are nearly identical: among learning disabled students as well, the use of recorded accommodations other than extra time and separate location was less common in New York than in Kentucky (Table 6).

Table 6

Comparison of Use of Accommodations, New York and Kentucky (Excluding Extra Time and Separate Location)

	All students with disabilities		Students with learning disabilities	
Number of accommodations	NY	KY	NY	KY
None	52.8	38.5	50.0	38.8
One	18.5	20.1	19.1	22.5
More than one	28.8	41.4	30.9	38.7

The second of these concerns can be addressed by focusing on accommodations that appear to be similar across the two states. Taken together, two of the classifications of accommodations used in New York, "directions read/clarified" and "test read to student," may correspond fairly well to two categories used in Kentucky, "oral presentation" and "paraphrasing." If so, this again suggests less frequent use of accommodations in New York than in Kentucky. In New York, 42% of students with disabilities had either "directions read/clarified" or "test read to student" (or both). In Kentucky, 58% of students were given either "oral presentation" or "paraphrasing."

Performance of Students With Disabilities

Interpreting the performance of students with disabilities is complicated by the uncontrolled use of accommodations. Decisions about the use of accommodations for specific students are made locally and are not clearly circumscribed, and data about the characteristics of students receiving different types of accommodations are very limited. One might expect that in general, students with more severe disability-related deficits in performance might be given more extensive accommodations, but these accommodations might then offset their tendency to score low. The actual impact of accommodations on performance can only be ascertained by experiments in which the use of specific accommodations for students with specific disabilities is varied systematically. Very few such studies have been conducted, and opportunities to conduct them are very limited.

Even in the absence of experimental data that could isolate the effects of accommodations from the effects of student characteristics, however, simple data on the performance of students with disabilities can provide clues to the quality of

measurement. For example, they can indicate the level of test difficulty for students with disabilities and can identify patterns in performance—such as mean differences between students with and without disabilities, differences in performance associated with accommodations, and anomalous item-level performance (e.g., differential item functioning, or DIF)—that suggest hypotheses and point to needed additional investigation.

The first of the following sections presents completion rates as a function of disability and accommodation. The second section discusses the overall performance of students with disabilities, separately by the type of accommodations they received. The final section explores whether performance varied as a function of the characteristics of forms and items.

Completion Rates

Completion rates provide an indication of whether the test is differentially speeded for students with and without disabilities. Low completion rates can reflect other aspects of difficulty as well; for example, students may fail to complete an assessment if they have become sufficiently demoralized.

Completion rates on the Regents English test, however, showed few differences among groups. Nearly all students completed the MC section of the form they were administered, and roughly three fourths completed both of the OR items they were administered (Table 7). Completion rates for students with and without disabilities were nearly identical. The one group difference that might be noteworthy is the higher percentage of unaccommodated students with disabilities who failed to complete either of the two OR items administered to them. Only about 12% of this group failed to complete either item, however, and the difference among groups could easily reflect only chance variation. Students who received accommodations on both sections had slightly lower completion rates than unaccommodated students despite the fact that more than half of the accommodated students received extended time. Although accommodations might be expected to increase the rate of completion, it is likely that the students who received accommodations had more severe disabilities and weaker prior achievement than those who didn't, and perhaps accommodations were not quite sufficient to offset that lower performance in terms of completion. Unfortunately, we only have descriptive data from a single field test administration and therefore cannot examine these selection effects.

Table 7

Completion Rates by Disability Status and Accommodations

Student group	Percent completed MC section	Percent completed both OR items	Percent completed one OR item	Percent completed no OR items
Non-disabled	97.0	78.7	17.7	3.6
All students with disabilities	95.9	73.5	18.5	8.0
Students with disabilities, unaccommodated	96.2	69.8	18.2	11.9
Students with disabilities, accommodated	95.8	75.0	18.6	6.4
Students with disabilities, accommodations including time extension	96.1	73.7	18.5	7.8
Students with disabilities, accommodations other than time extension	94.8	79.2	18.8	2.1

Because of the particular relevance of time extensions to completion, we calculated completion rates separately for disabled students with time extensions and for those with other accommodations but without time extensions. The completion rates for these two groups were very similar to the rates for all disabled students receiving accommodations, as indicated in Table 7. Those with time extensions had a trivially higher completion rate on the MC items and a trivially lower completion rate on the OR items than did disabled students with other accommodations (Table 7). The results for students with learning disabilities (not displayed) were, again, very similar to those shown for all students with disabilities.

Overall Performance Correlates of Accommodations

Except for completion rates, which are necessarily computed for the entire sample, performance was analyzed only for students who had informative scores on both the MC and the OR portions of the assessment. Students who did not respond to either section were dropped, as were students who had responses for only one of the two OR items administered to them. Of students who had responses to the MC portion of the assessment, a very small number, fewer than 2%, failed to complete all the items administered to them; these few students were not dropped. The exclusion of students with incomplete data caused the loss of a large number of cases: 21% of

the sample of students without disabilities and roughly 30% of the disabled sample.[3] This is detailed in Table 8. More students were missing OR scores than MC scores.

The forms administered in the Regents English field test were not comparable, and the number of students with disabilities administered each form was too small for most analysis. Therefore, a method was needed to make performance sufficiently comparable to allow pooling of data across forms. The MC items were scored as correct or incorrect, and each OR item was scored on a 6-point scale. In the following tables, the MC percent-correct scores have been standardized to a mean of 0 and a standard deviation of 1 within the students without disabilities sample. Therefore, the mean scores for students with disabilities are also the mean differences between students with and without disabilities, expressed as a fraction of the standard deviation in the population without disabilities. OR scores are the sum of the two-item scores (and therefore range from 0 to 12). These were also standardized to have a mean of 0 and a standard deviation of 1 for students without disabilities and are therefore in that sense comparable to the standardized MC scores.

On average, the Regents English field test was difficult for students with disabilities, regardless of whether they received accommodations. Students with disabilities received average scores that were approximately three quarters of a standard deviation lower than those of students without disabilities (Table 9). Results for students with learning disabilities were largely similar, although with no accommodations, the performance of students with learning disabilities was somewhat weaker than that of all students with disabilities (Table 10).

Table 8

Sample Loss From Incomplete Test Data

Student group	Total *N*	*N* with both MC and OR score	Percent lost
Non-disabled	8187	6429	21
Students with disabilities, unaccommodated	159	110	31
Students with disabilities, accommodated	404	300	26

[3] In the field test, papers were coded 7 if they were "unscorable, off-assessment, or straight copying from the text" (K. Kolanowski, personal communication, July 6, 2000). These cases were treated as missing in this analysis and account for roughly half of the cases of students with disabilities lost because of incomplete test data.

Table 9

Performance of All Students With Disabilities by Accommodations Category

	MC	OR	*N*
No accommodations	-0.65	-0.74	110
Any accommodations	-0.95	-0.84	300
Time extension only	-0.93	-0.79	46
Separate location only	-0.77	-0.92	21
Time extension and separate location	-0.71	-0.63	38
Time extension, separate location, and test read	-1.31	-1.09	20
Time extension, separate location, and directions	-1.02	-1.00	23
Time extension, separate location, test read, and directions	-0.77	-0.67	31

Table 10

Performance of Students With Learning Disabilities by Accommodations Category

	MC	OR	*N*
No accommodations	-0.84	-0.90	68
Any accommodations	-0.94	-0.84	255
Time extension only	-0.98	-0.77	41
Separate location only	-0.79	-1.00	18
Time extension and separate location	-0.76	-0.68	34
Time extension, separate location, and test read	-1.21	-1.15	18
Time extension, separate location, and directions	-1.04	-1.01	21
Time extension, separate location, test read, and directions	-0.72	-0.56	27

The mean differences between students without disabilities and students with disabilities varied from .65 to 1.31 standard deviations, depending on accommodations and item format (Table 9). The corresponding differences between students with learning disabilities and students without disabilities showed about the same range, from .56 to 1.21 standard deviations. These variations among

accommodations groups are difficult to interpret because of selection (that is, other differences among students receiving different accommodations) and the small number of students in each accommodation condition. However, if one assumes that selection should affect MC and OR questions approximately equivalently, there are two plausible explanations for these score variations. First, the effect of accommodations may differ between the OR and MC parts of the test. Second, in the case of some students, one or more of the indicated accommodations may have been used with only one part of the test.

Extended time appears to give more of a boost to performance on OR items than on MC items. Most groups of accommodated students listed in Tables 9 and 10 scored lower—often substantially lower—than did students with no accommodations, presumably because accommodations are more likely to be offered to students with more severe disabilities and lower performance levels. However, in the groups receiving extended time either alone or in combination with other accommodations, the gap between accommodated and unaccommodated students was larger on MC items than on OR items. This held true for both students with learning disabilities and all students with disabilities.

To illustrate this pattern, consider students with disabilities receiving no accommodations or a time extension alone. On the MC items, the group with no accommodations had a mean of -.65 (.65 standard deviation below the mean of non-disabled students), while the group receiving time extensions alone had a mean of -.93 (Table 9), a drop of .28 standard deviation. In contrast, the group getting a time extension alone scored only .05 standard deviation lower than the group with no accommodations on the OR items. Thus, the additional benefit of the accommodation for OR performance was .23 standard deviation. The four other combinations in Table 9 that include time extension showed additional benefits for OR performance ranging from .11 to .31 standard deviation. Across all accommodations conditions, additional time was associated with an increase of .13 standard deviation on the OR portion of the test but essentially no change on the MC portion. The increase in OR scores associated with extended time, however, was not statistically significant, a function of the small numbers of students in each group.[4]

[4] These estimates are based on ordinary least squares regressions in which standardized scores (MC and OR separately) were regressed on four dummy variables indicating the presence or absence of time extension, separate location, reading of directions, and reading of the test.

Without experimental control of accommodations or additional information about the performance of students administered the Regents English test, one cannot in general determine the validity of scores obtained by students with disabilities. One can, however, look for patterns in the scores that are implausible, such as very high means or unreasonably large differences between groups receiving different types of accommodations. For example, in 1995, Kentucky elementary school students with learning disabilities or mild mental retardation who received certain combinations of accommodations including dictation (use of a scribe) had implausibly high scores. Learning disabled students receiving these accommodations had substantially higher scores than did students without disabilities, and mentally retarded students had nearly average scores (Koretz, 1997). These patterns had disappeared two years later (Koretz & Hamilton, 1999).

None of the means for students with disabilities in the Regents English field test, however, appear implausible on their face. None of the means in Table 9 are implausibly high. The differences among accommodations conditions are also modest, even though the very small size of some of the accommodations groups increases the probability that unreasonable results would have occurred by chance.

Performance Correlates of Form and Item Characteristics

In most instances, the performance of students with disabilities is described here relative to that of students without disabilities. Because the forms administered in the field test differed in length, however, it is important to compare the difficulty of the forms before describing within-form differences in performance between students with and without disabilities. The difficulty of forms is shown by the raw scores of students without disabilities; that is, the percent of MC items answered correctly and the sum of the two OR scores without any standardization.

Three characteristics of forms were examined. One was their length, which is the number of MC items they included. The other factors pertained to the types of prompts used to elicit extended responses. The Regents English test includes two unusual prompts, and these were singled out to receive special attention. One of these is the prompt presented orally rather than in writing (Part I, "Listening and Writing for Information and Understanding"). The other is the prompt requiring students to write about two literary works they've previously read (Part IV, "Reading and Writing for Critical Analysis"). As noted earlier, the latter prompt presents students with a statement or quotation about literature. They are asked to

explain it, state their opinion about it, select two literary works they have read that support their opinion, and discuss specific elements of these works in developing support for their arguments.

Form length. Difficulty can be gauged in numerous ways. Here we look at the difficulty of forms by considering completion rates, raw mean performance, and the difficulty of forms for students with disabilities relative to students without disabilities.

Although one might expect the length of the forms to have an impact on completion rates, there was no strong relationship between the number of MC items included in the Regents English forms and the rate of completion of the MC section of the test. There was a tendency for completion rates to decrease as the number of MC items increased, but the differences were small and the pattern was inconsistent across the three groups of students (Table 11). We also looked at completion rates for OR items because as the number of MC items increases, the amount of time available to complete the OR items may decrease. Students could compensate for the larger number of MC items on some forms by allocating less time to the OR items, and that might be reflected in the percentages of students who completed both of the OR items administered to them. The completion rate for OR items, however, also showed no consistent differences among forms or groups (Table 11).

The forms of different lengths (that is, with different numbers of MC items) varied somewhat in average performance levels, but there was not a consistent relationship between difficulty and length, and some of the differences were small. The 20-item forms were on average the most difficult, showing the lowest mean scores on both the MC and OR components, and the 6-item forms were the easiest (Table 12). The 16-item forms, however, were nearly as easy as the 6-item forms, and the 10-item forms were nearly as difficult as the 20-item forms. This pattern may be due to the presence of a read-aloud passage on the 6- and 16-item forms, discussed later.

To put these differences in perspective, the standard deviation of the total OR score was 1.9. Thus the six differences between the OR means of different-length forms ranged from .05 to .39 standard deviation. Similarly, the differences between MC means ranged from .03 to .33 standard deviation.

Table 11

Completion Rates on MC Section by Number of MC Items Administered, Disability, and Accommodations

Number of MC items administered	Disability and accommodation status	Percent completed MC section	*N*	Percent completed both OR items	*N*
6	Non-disabled	99.1	1606	82.8	1624
	Disability, no accommodations	100	36	81.1	37
	Disability, accommodations	100	65	84.6	65
10	Non-disabled	98.3	2541	75.6	2547
	Disability, no accommodations	95.4	65	53.8	65
	Disability, accommodations	96.7	121	65.9	123
16	Non-disabled	95.9	2632	80.7	2637
	Disability, no accommodations	92.9	28	82.1	28
	Disability, accommodations	97.3	150	82.1	151
20	Non-disabled	94.0	1376	76.1	1379
	Disability, no accommodations	96.6	29	79.3	29
	Disability, accommodations	85.9	64	66.2	65

Table 12

Performance of Non-Disabled Students by Number of MC Items (Raw Scores: MC Percent Correct and OR Total)

Number of MC items administered	MC	OR	*N*
6	80.4	6.2	1335
10	75.6	5.6	1920
16	79.8	6.0	2126
20	74.4	5.5	1048

The length of forms did, however, appear to affect the relative difficulty of MC items for students with disabilities. Students with disabilities scored .77 standard deviation and .58 standard deviation below students without disabilities on the MC sections of the 6- and 10-item forms, respectively (Table 13). However, students with disabilities scored more than a full standard deviation below students without disabilities on the MC sections of the longer forms. In contrast, the relative performance of students with disabilities on the OR sections of the forms was unaffected by form length. Small sample sizes make it difficult to contrast students with and without accommodations, but there is no apparent consistent difference between them in this respect; both show a decline in MC performance but not in OR performance as forms are lengthened (Table 14).

Inclusion of a read-aloud passage. Half of the forms included a read-aloud passage. These also included one other item requiring an extended response. The read-aloud passage was always in the first of the two parts of the form and included 1 OR and 6 MC items.

Table 13

Performance of Non-Disabled Students by Number of MC Items in Form (Raw Scores: MC Percent Correct and OR Total)

Number of MC items administered	MC	OR	*N*
6	-0.77	-0.86	84
10	-0.58	-0.70	115
16	-1.05	-0.87	146
20	-1.11	-0.83	65

Table 14

Performance by Number of MC Items in Form (Standardized Scores)

	No accommodations		Accommodations		
Number of MC items	MC	OR	MC	OR	*N*
6	-0.52	-0.78	-0.90	-0.91	29
10	-0.41	-0.71	-0.65	-0.69	80
16	-1.01	-0.61	-1.05	-0.92	123
20	-0.83	-0.84	-1.27	-0.82	42

Among students without disabilities, the forms with read-aloud items were easier than others. On both the OR and MC parts of the assessment, these students received higher average scores on forms with read-aloud passages (Table 15).

The performance of students with disabilities, however, fell modestly farther behind that of students without disabilities on forms with read-aloud prompts compared with other forms. That is, while read-aloud forms were easier for students without disabilities, the relative difficulty of these forms, comparing students with and without disabilities, was somewhat greater. For example, students with disabilities who received accommodations obtained an average standardized OR score of -.74 on forms without read-aloud prompts and a score of -.91 on forms that included a read-aloud prompt, a difference of .17 standard deviation (Table 16). The MC portion of the test showed differences of roughly this magnitude regardless of accommodation. In the case of OR items, however, there were two exceptions: students with disabilities who received no accommodation or accommodations without time extension performed trivially higher on the read-aloud forms than on other forms relative to students without disabilities. These patterns suggest that accommodations, apart from those that include no time extension, may boost OR performance for items presented in writing but not for items presented orally.

A comparison of the two halves of each form suggests that the read-aloud items themselves, rather than some other aspect of these forms, account for the patterns described. Table 17 shows the differences in performance (raw scores) between read-aloud and other forms, separately for Part A and Part B of the forms. The read-aloud item was always Part A of the form. Thus, the top (Part A) panel of Table 17 contrasts read-aloud to other items, while the bottom (Part B) panel contrast two items that were not read aloud. Read-aloud items in Part A generated fewer low scores and higher mean scores for all three groups: non-disabled, disabled with no accommodations, and disabled with accommodations. (The mean was higher by a smaller amount [.36 standard deviation] for students with disabilities who were accommodated than for others—another indication of the greater relative difficulty of read-aloud items for students with disabilities.) In contrast, Part B, which never included a read-aloud item, showed no consistent difference between forms with and without read-aloud items.

Table 15

Performance of Non-Disabled Students on Forms With and Without Read-Aloud Passages (Raw Scores: MC Percent Correct and OR Total)

Condition	MC	OR	*N*
No read-aloud	75.2	5.5	2968
Read-aloud	80.1	6.1	3461

Table 16

Performance of Students With Disabilities on Forms With and Without Read-Aloud Prompts, by Accommodation (Standardized Scores)

Student group	No read-aloud			Read-aloud			Difference	
	MC	OR	*N*	MC	OR	*N*	MC	OR
No accommodations	-0.58	-0.76	58	-0.74	-0.71	52	-0.16	0.05
With accommodations								
Any accommodations	-0.86	-0.74	122	-1.00	-0.91	178	-0.14	-0.17
Test not read	-0.80	-0.69	69	-0.94	-0.85	111	-0.14	-0.16
Test read	-0.95	-0.80	53	-1.12	-1.02	67	-0.17	-0.22
No time extension	-0.88	-1.03	26	-1.09	-0.96	50	-0.21	0.07
Time extension	-0.86	-0.66	96	-0.97	-0.90	128	-0.11	-0.24

Table 17

Performance on OR Items by Position, Disability, and Accommodation (Raw Scores)

	Non-disabled	Disabled: no accommodation	Disabled: any accommodation
Part A			
Percent scored 1	-5.4	-10.5	-13.5
Percent scored 1 or 2	-18.5	-30.5	-13.3
Mean	0.63	0.52	0.36
Part B			
Percent scored 1	-0.7	-3.5	6.2
Percent scored 1 or 2	4.7	-5.1	11.8
Mean	-0.10	0.09	-0.22

Inclusion of a prompt requiring students to write about previously read works. Another type of prompt, Part IV, "Reading and Writing for Critical Analysis," presents students with a statement or quotation about literature. They are asked to explain it, state their opinion about it, select two literary works they have read that support their opinion, and discuss specific elements of these works in developing support for their arguments. For brevity, forms that include this prompt will be referred to as "previously read" forms.

In contrast to the read-aloud forms, the previously read forms were as difficult as other forms for non-disabled students. On both the MC and OR parts of the assessment, raw scores on these forms were nearly identical to those on other forms (Table 18).

The previously read forms, however, did differ from others in terms of their relative difficulty for students with disabilities. The OR scores of students with disabilities, expressed relative to the performance of students without disabilities, were roughly the same on the two types of forms, regardless of whether accommodations were provided (Table 19). On the MC portion of the assessment, however, the two types of forms differed greatly in relative difficulty. The MC portion of the previously read forms was much easier, relatively, than the MC portion of other forms for students with disabilities, particularly for those who

Table 18

Performance of Non-Disabled Students on Previously Read and Other Forms (Raw Scores: MC Percent Correct and OR Total)

Condition	MC	OR	*N*
Previously read	78.0	5.8	3174
Other	77.6	5.8	3255

Table 19

Comparison of Forms With Literature-Based Essay and Those Without (Standardized Scores)

	Previously read			Other		
Student group	MC	OR	*N*	MC	OR	*N*
Disability, no accommodations	-0.46	-0.74	64	-0.92	-0.73	46
Disability, accommodations	-0.75	-0.78	135	-1.11	-0.89	165

received no accommodations (.46 versus .92 standard deviation; see Table 19). The reason for this difference cannot be ascertained from these data, but the length of the forms may have contributed. The previously read forms have only one MC section, while other forms have two.

Item-Level Analyses

The small sample sizes precluded some types of item analysis, such as formal tests of differential item functioning (DIF). However, it was possible to explore descriptive information about several aspects of item functioning: difficulty, discrimination, and differences in item functioning across groups.

Item Difficulty

Although the Regents English test was much harder, on average, for students with disabilities than for other students, performance on MC items taken individually did not suggest that these items were so difficult as to be uninformative for students with disabilities. Across all items, the mean p values for disabled students with and without accommodations were substantially lower than the p value for students without disabilities (Table 20). Few items, however, showed very low p values for students with disabilities. Only 8% of items showed p values below .3 for disabled students without accommodations, and only 6% did for students with accommodations.

OR items, however, present a very different picture, with indications that some items may have been excessively difficult for students with disabilities. Three measures were used to gauge the difficulty of individual OR items. The first was the percentage of uninformative responses. These included responses coded as 7, indicating "unscorable, off-assessment; or straight copying from the text," or coded as 0, indicating submission of a blank paper. The second measure was the percentage of responses scored as a 1, after omitting those coded 7 or 0. The rubric

Table 20

Difficulty of MC Items, by Group

	Mean p value	Percent p values < .3
Non-disabled	0.78	0.01
Disabled, no accommodation	0.65	0.08
Disabled, any accommodation	0.60	0.06

categorized scores of 1 as responses that "provide minimal or no evidence of textual understanding," "show no focus or organization," "use language that is incoherent or inappropriate," and "may be illegible or not recognizable as English." The third measure was the percentage of responses scored either 1 or 2. Scores of 2 are characterized by the rubric as responses that "convey a confused or inaccurate understanding of the text," "are incomplete or largely undeveloped," "lack an appropriate focus but show some organization," "use language that is imprecise or unsuitable for the audience or purpose," and "exhibit frequent errors that make comprehension difficult."

Among students without disabilities, the percentage of 0 or 7 responses was negligible in the case of several items and reached a maximum of roughly 10% in the case of five items (Figure 1).[5] In contrast, the percentage of students with disabilities that scored 0 or 7 was 10% or higher for 11 of the 28 OR items and reached a maximum of about 30%.

When students scoring 0 or 7 were excluded, the percentages scoring 1 showed an even more striking contrast between disabled and non-disabled students. Among non-disabled students, these percentages were generally below 15% and reached a maximum of roughly 20% (Figure 2). In contrast, the percentage of students with disabilities scoring 1 exceeded 20% for all but 8 items. This percentage exceeded one third for 9 of 28 items and exceeded 40% for 6 items. When more than 40% of a group submit responses that raters characterize as "illegible or not recognizable as English" and the like, it seems likely that the item is too difficult for the group in question. Because some of the OR items did not have high percentages of students with disabilities scoring 1, the requirement of writing as such could not explain the excessive difficulty. The requirement of writing may interact with content or other demands of the tasks, however, to make some OR items too hard for some students with disabilities.

A large number of students, both with and without disabilities, scored 2 on most items, and the contrast between disabled and non-disabled students is less extreme when the percentages scoring either 1 or 2 are compared (Figure 3). Nonetheless, the poor performance of students with disabilities on some items is striking. The percentage of students scoring either 1 or 2 reaches a maximum of

[5] This figure was drawn by sorting items in terms of the percentage of students with disabilities who scored 0 or 7. Because the corresponding percentages for non-disabled students are not perfectly correlated with these percentages, the line for students with disabilities appears erratic.

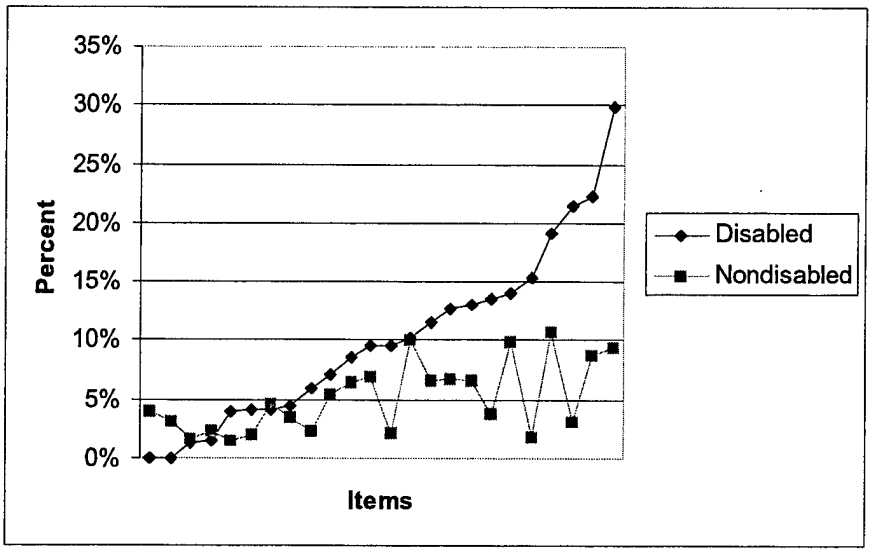

Figure 1. Percent of OR responses scored 0 or 7, by group (sorted by percentage for students with disabilities).

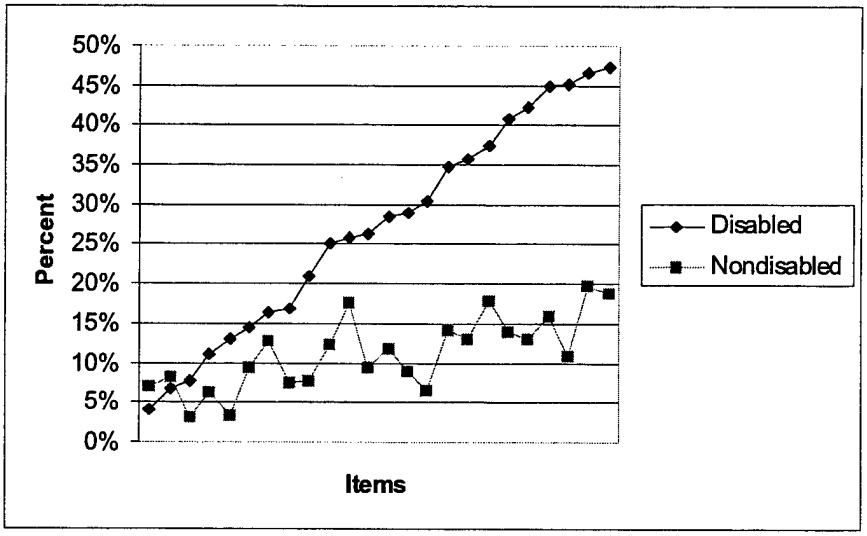

Figure 2. Percent of OR responses scored 1, by group (sorted by percentage for students with disabilities).

roughly 50% for students without disabilities. For students with disabilities, on the other hand, it reaches a mean of 93%, and it reaches or exceeds 75% for 7 of the 28 items.

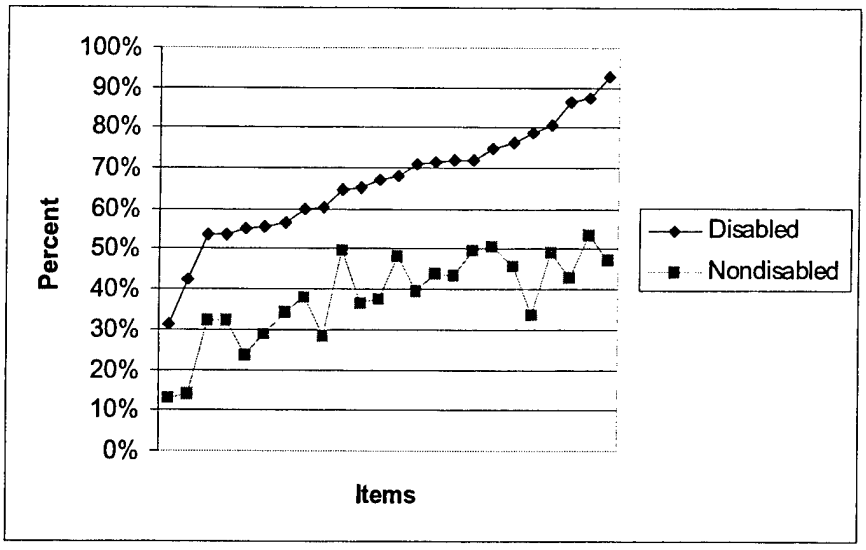

Figure 3. Percent of OR responses scored 1 or 2, by group (sorted by percentage for students with disabilities).

Item Discrimination

After the first (1995) field test of greater inclusion of students with disabilities on the National Assessment of Educational Progress, Anderson, Jenkins, and Miller (n.d.) found that the correlations between items and total scores often were lower for students with disabilities than for other students. This indicated that the assessment was less discriminating for students with disabilities. Koretz (1997) and Koretz and Hamilton (1999), in contrast, did not find any difference in discrimination in the Kentucky KIRIS assessment between non-disabled students and either all students with disabilities or students with learning disabilities.

Because of the small number of students with disabilities in the Regents English field test, particularly those who received no accommodations, item-test correlations would be expected to vary markedly among the student groups simply because of sampling error. Accordingly, differences between groups in particular item-test correlations would not be meaningful. One can, however, compare the distributions of these correlations.

In the case of MC items, the distributions of item-test correlations suggest that discrimination is roughly the same in all three groups: students without disabilities, students with disabilities who received no accommodations, and students with disabilities who received accommodations. The means and medians of the item-test correlations are quite similar across the three groups, and the correlations for

students with disabilities but no accommodations are slightly higher than those for the other groups (Table 21). The correlations are considerably more variable among students with disabilities, particularly among the group without accommodations (see Figure 4). These differences in variability, however, are expected, given differences in sample size.[6]

Table 21

Mean and Median of Item-Test Correlations, MC Items, by Group (Point-Biserial Correlations)

	Mean	Median
Non-disabled	.38	.37
Disabled, no accommodations	.41	.46
Disabled, any accommodation	.35	.36

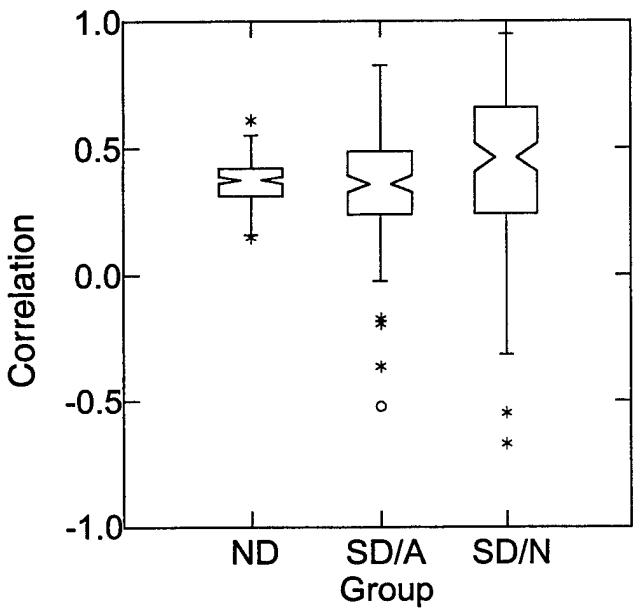

Figure 4. Distributions of item-test correlations, MC items, by group. Note that the center of the notch on each box plot is the median, and the notch itself spans a 95% confidence interval around the median. As in a conventional box plot, the vertical distance between the ends of the boxes represents the range from the 25th to the 75th percentile. The number of students in each group is not represented in the plot, although it influences both the spread of the plot and the size of the confidence interval shown by the notch.

[6] The small size of these correlations stems in part from the fact that they are point-biserial correlations, which are lower than Pearson correlations—and bounded at a value below 1—because the dichotomous distribution of item scores cannot fully match the continuous distribution of test scores.

The distributions of item-test correlations for OR items also do not show clear evidence of lower discrimination for students with disabilities. They do suggest lower discrimination for students with disabilities who received no accommodations. Because this group was very small, this pattern should not be given much weight without replication. The mean correlations are slightly lower for both groups of students with disabilities, and the median is lower as well for students without accommodations (Table 22). The stronger sign of lower discrimination, however, appears when all of the distributions are viewed graphically (Figure 5). The entire distribution of correlations for disabled students without accommodations is shifted downward relative to that for students without disabilities. As indicated by the very wide notch on the box plot for students without accommodations, however, the small sample leaves little confidence in the

Table 22
Mean and Median of Item-Test Correlations, OR Items, by Group

	Mean	Median
Non-disabled	.73	.73
Disabled, no accommodations	.63	.65
Disabled, any accommodation	.68	.72

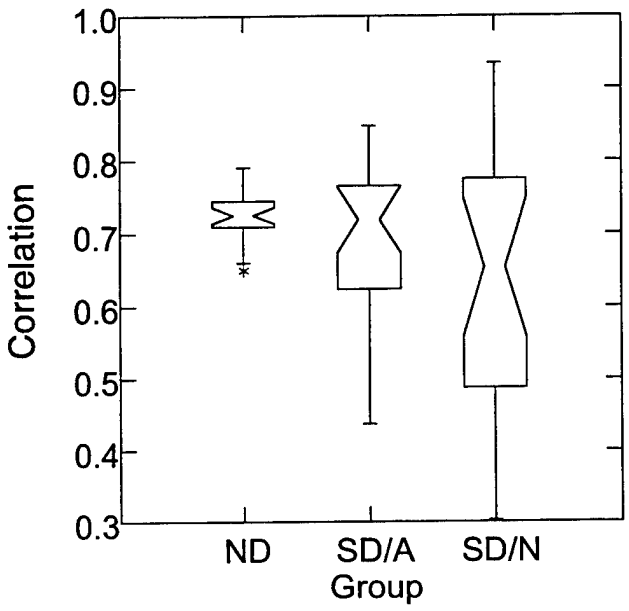

Figure 5. Distributions of item-test correlations, OR items, by group.

distribution for that group. Only more extensive data could determine whether these differences are important or only a matter of sampling.

Differences in Item Functioning Across Groups

Differences in the functioning of test items across groups generally is called differential item functioning, or DIF; we avoid this term here because the small sample made formal tests of DIF impractical.

In lieu of formal tests of DIF, we examined the distributions of group differences on individual OR items to identify items that showed unusually large or unusually small differences between students with and without disabilities. These comparisons used raw differences between students without disabilities and disabled students who received accommodations. This is strictly comparable to most tests of DIF only under certain restrictive conditions, but it does provide a first look at possible differential difficulty for students with disabilities.[7]

The OR items varied markedly in terms of the mean differences between students without disabilities and students with disabilities who received accommodations. The smallest mean difference was 0.42, while the largest was 1.12. Although the mean differences were distributed across the entire range, one cluster of items showed mean differences of roughly 0.5, while another showed differences of roughly 1.0 (Figure 6).

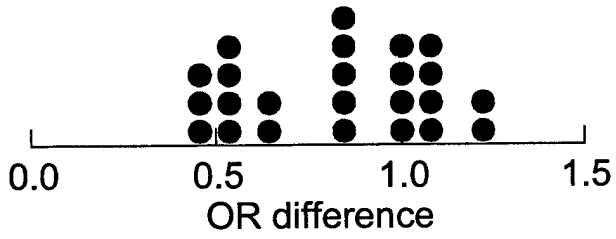

Figure 6. Group differences in difficulty, OR items (mean difference between non-disabled and accommodated; each symbol is one item).

[7] Most tests of DIF examine whether students in different groups but with equivalent proficiency overall score differently on a specific item. That is, they look for a group difference in performance on a particular item, holding constant total score. The approach used here, which compares scores across groups without holding total scores constant, is comparable to these tests of DIF only if the correlation between item performance and total score is similar across both groups and items.

The size of mean differences was related to the type of item. Five of the six "previously read" items (that is, items that required students to write about literature they had read previously) showed small mean differences. Five of the six read-aloud items showed large differences. It is not clear whether these patterns stem from the small sample size, and the reasons for them cannot be ascertained without additional data. These findings do, however, raise concerns about possible interactions between item format and the relative performance of students with disabilities.

Discussion

The limitations of the data generated by the Regents English field test rule out drawing many firm conclusions about the suitability of the assessment for students with disabilities and about the performance of these students. Only very large differences in performance can be statistically significant because of the small samples, and other limitations of the data cloud the causal interpretation of even large differences. Nonetheless, the patterns found here are informative and have implications for policy and research.

Patterns in the Findings

In this section, we summarize and integrate the findings. We discuss their implications in the following section.

How inclusive was the field test? Comparisons of the percentages of students with disabilities in the tested sample and in New York's population of students suggest that the field test excluded a sizable percentage of students with disabilities, particularly students with disabilities other than learning disabilities. At the very least, the tested sample included fewer such students than does the state's population of students.

Because of the nature of the field test, the low disabled-student participation rate could stem from any number of factors, and the implications for the inclusiveness of the operational Regents English assessments is uncertain. For example, it may be that the sampling for the field test resulted in a sample of schools that included an atypically small number of students with disabilities. It also is possible that schools will work more diligently to include students with disabilities in operational assessments than in field tests. The possibility remains, however, that these findings presage substantial non-participation of students with disabilities in

the new Regents examinations once they become operational, particularly given that the requirement to take the Regents examinations is new for many students with disabilities.

How were accommodations used? Accommodations were used liberally; nearly three fourths of participating students with disabilities were given one or more accommodation. Time extension and separate location were both provided to more than half of tested students with disabilities, and "test read" and "directions read" were both provided to about 30%. It is not clear, however, whether the use of accommodations was either greater than intended by the New York guidelines or more extensive than warranted by the goal of maximizing validity.

Accommodations other than time extension and separate location were used somewhat less frequently in the Regents English field test than in the operational 11th-grade assessment in Kentucky, one of the few states that can provide information on the use of accommodations in an inclusive assessment system. The cause of this difference, however, is unclear. For example, it could stem from differences in the characteristics of the assessment, state guidelines, the participating samples, or the level of stakes.

How difficult was the assessment for students with disabilities? Several of the findings shed light on the difficulty of the assessment for students with disabilities: completion rates, overall mean performance, and information on item-level difficulty. (Differences in difficulty across forms of different types are discussed later.)

The picture painted by these diverse measures is mixed. Given the extensive use of accommodations, however, the possibility remains that some measures understate the difficulty of the assessment for students with disabilities.

Two measures, completion rates and p values for MC items, did not suggest that the test was particularly difficult for students with disabilities. Completion rates showed little consistent difference among groups of students, and few MC items had very low p values for students with disabilities.

The mean scores of students with disabilities, however, were markedly lower than those of non-disabled students. When students were placed in groups on the basis of the accommodations they received, the mean scores of the groups with disabilities ranged from roughly two thirds to one and a third standard deviation below the mean of non-disabled students. To put these differences in perspective,

they are roughly comparable to the mean differences between Black and White students shown on a variety of large-scale assessments of achievement (e.g., Hedges & Nowell, 1998).

Moreover, the high percentage of students with disabilities who provided either unscorable or extremely weak responses to many of the OR items also suggests that the OR portions of the test are very difficult for some students with disabilities. Particularly in light of the extensive use of accommodations, these findings suggest that many of the OR items are simply beyond the reach of many students with disabilities.

How did performance vary with accommodations? Overall, performance was roughly similar for disabled students with and without accommodations but varied substantially among groups that received different accommodations. Extra time appeared to raise performance on OR items more than on MC items, although the difference was modest and was not statistically significant. This difference, if it is real and not random variation, could indicate a greater need for accommodations in OR tests, or it could represent excessive effects of some accommodations on OR performance (see Koretz, 1997). Additional research exploring this could be fruitful because if extended time or other accommodations do have a larger impact on one format than on another, the relative scores of students with disabilities will be sensitive to the weight given to each of the formats in operational assessments.

None of the mean scores of the groups receiving different accommodations were implausible on their face, but that is not sufficient basis for accepting as valid the scores of students with accommodations, that is, to decide that the effects of accommodations are appropriate and increase validity. First, the uncontrolled assignment of accommodations means that students with and without accommodations may differ in important respects. For example, it may be the case that students with accommodations would have been lower performing on average than those without accommodations if no accommodations had been offered. In this case, accommodations would be offsetting the otherwise lower performance of students who received them. In the absence of additional descriptive information about participating students, there is no way to explore possible selection effects of this sort. Second, the small sample sizes make large random variations in the means of groups receiving specific accommodations likely.

Did performance vary among types of tasks? In terms of mean scores, neither MC nor OR items placed students with disabilities at a particular disadvantage on the Regents English test. Although some differences appeared for groups receiving specific accommodations, students with disabilities overall performed roughly comparably on the two formats. Whether that would remain true if the sample were more representative or if the use of accommodations were different remains unclear. As noted, however, OR items did pose a far greater problem for students with disabilities in terms of the percentage of items that appeared to be too difficult for an appreciable number of students.

Several types of analysis suggested that the type of OR prompt did have a bearing on the relative performance of students with disabilities. In particular, forms with read-aloud prompts, which were less difficult than others for students without disabilities, were relatively more difficult for students with disabilities. This appears to have been a result of the read-aloud prompt in those forms. There was also some evidence that prompts that required students to write about previously read literature were relatively easier for students with disabilities, but this did not create a clear difference in performance on the forms that included them.

Implications

The findings discussed here suggest the need for additional information that would help inform policy and underscore several issues now confronting policy-makers who are deciding how best to assess students with disabilities.

Both the limitations of the field test data and the unavoidable differences (e.g., in motivation) between field tests and operational, high-stakes assessments suggest the importance of monitoring both the rate of exclusion of students with disabilities and the use of accommodations. Because prevalence data are reported by broad age range rather than grade, and the new Regents tests can be taken by students within a range of grades and ages, it will be necessary to monitor examinations for several years in order to estimate the participation rates of students with disabilities. In addition, to judge these participation rates, it will be important to set a target for them based on the characteristics and uses of the examinations and the nature of the alternatives open to students.

The results here also suggest the importance of additional exploration to help judge the appropriateness of the current uses of accommodations. Collecting additional descriptive data about students with disabilities in the context of an

operational form of the assessment might be a low-cost and relatively non-intrusive way to learn more about the uses of accommodations. But it would not be sufficient to ascertain the effects of accommodations on the level or validity of scores. Determining the validity of scores with various kinds of accommodations will likely require experimental data as well as richer descriptive data. There have been very few experimental studies of accommodations in K-12 education, and some of the few such studies have examined narrow accommodations, such as putting answer bubbles for MC questions in the test booklet rather than on a separate answer sheet. One reason for the dearth of experimental studies is that some see them as politically difficult because they would necessarily entail denying some students accommodations that they might otherwise receive. This difficulty could be lessened in several ways, for example, by experimenting only in the context of field tests and by comparing only combinations of accommodations that mirror likely guidelines for their actual use without a "no accommodations" condition. Absent this additional research, policymakers will have only a weak basis for decisions about how best to assess students with disabilities.

The results presented here also suggest the importance of additional exploration of possible differences in the difficulty of different types of items and forms used in the Regents English test for students with special needs, including both students with disabilities and English-language learners. Collection of simple descriptive information in conjunction with operational assessments would be sufficient to allow some useful investigation of these issues.

The difficulty of the Regents English test for some students with disabilities raises several important issues. The first is the quality of the performance information for students with disabilities. Tests typically provide less information at extreme values; that is, the information provided for students with extreme scores is less accurate. If the Regents English test will be used solely to provide an indication of whether students have reached the cut-score required for graduation, the accuracy of scores well below that is less important than the accuracy of scores around the cut-score. If performance on the Regents English serves other functions as well, for example, influencing placement, remediation, or course grades, then the accuracy of scores in the range achieved by many students with disabilities becomes more important.

The second issue is common to all standards-based reporting systems that indicate whether students have reached one or a few performance levels. The

Regents English test is given a score so students and teachers can differentiate levels of performance well below or above the state's cut-scores. However, to the extent that NYSED reports statistics such as the percentage of students passing each Regents examination—a traditional measure in New York—improvements in performance that fail to raise low-scoring groups to the cut-score, and improvements in the performance of groups already above the cut, go unrecognized. This in turn may lessen incentives to focus instructional efforts on these students.

A third issue raised by the difficulty of the assessment is the possible effect of requiring groups of students to take high-stakes tests that are very difficult for them. This issue pertains not only to students with disabilities but also to numerous other groups with low average scores. Although one positive effect over the moderate term may be the improvement of instruction for and learning within these groups, negative effects are possible as well. At least over the short term, failure rates may be very high unless the cut-score is set so that the overall failure rate is low. For example, assume that the initial failure rate (that is, the failure rate when each student first takes the Regents English test) is 30% for students without disabilities. Under those conditions, the failure rate for groups with a mean of -.65—the mean of the highest scoring group of students with disabilities shown earlier—would be expected to be roughly 55%. For students with a mean -1.33, the lowest of the means shown earlier, the expected initial failure rate would be about 78%. Assuming an overall mean of -1.0 standard deviation for students with disabilities would lead to an expected failure rate of 68%. Note that in this case as well, the true difficulty of the assessment may be understated if the use of accommodations is inappropriately generous.

Overly difficult tests may have other undesirable effects on students and also on teachers. Teachers, for example, may resort to instructional shortcuts or worse in an attempt to avoid high failure rates. For example, they may resort to inappropriate coaching or may unduly narrow the curriculum. Students may become demoralized by the prospect of facing a test on which they are likely to do poorly; some may even drop out of school.

The risks of administering tests that are overly difficult for some, however, must be weighed against the potential benefits of including as many students as possible in the system of standards and assessments. If groups of students are exempted, there is a risk that educators will not be held accountable for their

learning and that the students will be given inferior opportunities (e.g., National Research Council, 1997).

The questions raised here underscore the importance of evaluating the diverse effects of including students with disabilities—and other students with special needs—in assessment programs designed for the general education population. These effects are likely to vary, depending on the difficulty of standards, the types of assessments employed, and the types of accommodations and modifications offered. Only systematic research will reveal the ways of including these students in order to produce the greatest benefits while minimizing unintended effects.

References

Anderson, N. E., Jenkins, F. F., & Miller, K. E. (n.d.). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

Hedges, L. V., & Nowell, A. (1998). Black-white test score convergence since 1965. In C. Jencks & A. Phillips (Eds.), *The black white test score gap* (pp. 149-181). Washington, DC: Brookings.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., & Hamilton, L. (1999). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject* (CSE Tech. Rep. No. 499). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

National Research Council, Committee on Goals 2000 and the Inclusion of Students With Disabilities. (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

New York State Education Department. (1995). *Test access and modifications for individuals with disabilities.* Albany, NY: Office of Vocational and Educational Services for Individuals with Disabilities.

New York State Education Department. (n.d. a). *Sample for October 1998 field-testing for English Regents Examination.* Albany, NY: Author.

New York State Education Department. (n.d. b). *New Regents Comprehensive Examination in English: Summary of specifications.* Albany, NY: Author.

U.S. Department of Education. (1998). *Twentieth annual report to Congress on the implementation of the Individuals with Disabilities Education Act.* Washington, DC: Author. Available 16 April 2001: www.ed.gov/offices/OSERS/OSEP/OSEP98AnlRpt/